Laporan Praktikum Pertemuan 10

**Data Science Lanjut**

Machine Learning With Python For Begginer



Dibuat Oleh

Nama : Muhamad Faisal Halim

NIM : 19.240.0163

Kelas : -

Mata Kuliah : Data Science Lanjut

Mahasiswa Pertukaran Mahasiswa.

**Universitas Muhammadiyah Kalimantan Timur**

~ STMIK Widya Pratama Pekalongan

**Note**

Mohon maaf pak, pada materi sebelumnya masih terdapat beberapa tugas yang belum saya kumpulkan, yaitu pada module 3,4 dan 6 untuk laporan saya sertakapan pada link berikut [link\_1](https://drive.google.com/drive/folders/1qtaG1pREWQin83ZvYxPP4aF0Cs4iHXZp?usp=sharing) atau <https://bit.ly/laporan-tertinggal-halim0163>.

Semoga dengan saya mencantum laporan sebelumnya ini bisa mengisi kekosongan tugas laporan saya dan menjadi bahan pertimbangan tambahan untuk hal yang diperlukan dalam penilaian akhir matakuliah Data Science lanjut

**Materi dan Praktikum**

Title : **Machine Learning With Python For Begginer**

Teknologi machine learning (ML) adalah mesin yang dikembangkan untuk bisa belajar dengan sendirinya tanpa arahan dari penggunanya.

Dalam hal ini machine learning memiliki kemampuan untuk memperoleh data yang ada dengan perintah ia sendiri. ML juga dapat mempelajari data yang ada dan data yang ia peroleh sehingga bisa melakukan tugas tertentu. Tugas yang dapat dilakukan oleh ML pun sangat beragam, tergantung dari apa yang ia pelajari.

Istilah machine learning pertama kali dikemukakan oleh beberapa ilmuwan matematika seperti Adrien Marie Legendre, Thomas Bayes dan Andrey Markov pada tahun 1920-an dengan mengemukakan dasar-dasar machine learning dan konsepnya. Sejak saat itu ML banyak yang mengembangkan. Salah satu contoh dari penerapan ML yang cukup terkenal adalah Deep Blue yang dibuat oleh IBM pada tahun 1996.

Sebagai contoh lain, pernah buka folder spam di email kamu? Di sana ada banyak email promosi dan sebagainya yang dikirim secara acak ke email kita padahal tidak kita inginkan. Lalu bagaimana, perusahaan seperti Google, Yahoo, Hotmail, Microsoft dan perusahaan penyedia lainnya secara otomatis memfilter pesan spam ini? Tentunya dengan machine learning.

Terminologi Machine Learning

Dalam pembuatan model machine learning tentunya dibutuhkan data. Sekumpulan data yang digunakan dalam machine learning disebut DATASET, yang kemudian dibagi/di-split menjadi training dataset dan test dataset.

TRAINING DATASET digunakan untuk membuat/melatih model machine learning, sedangkan TEST DATASET digunakan untuk menguji performa/akurasi dari model yang telah dilatih/di-training.

Teknik atau pendekatan yang digunakan untuk membangun model disebut ALGORITHM seperti Decision Tree, K-NN, Linear Regression, Random Forest, dsb. dan Output atau hasil dari proses melatih algorithm dengan suatu dataset disebut MODEL.

Umumnya dataset disajikan dalam bentuk tabel yang terdiri dari baris dan kolom. Bagian Kolom adalah FEATURE atau VARIABEL data yang dianalisa, sedangkan bagian baris adalah DATA POINT/OBSERVATION/EXAMPLE.

Hal yang menjadi target prediksi atau hal yang akan diprediksi dalam machine learning disebut LABEL/CLASS/TARGET. Dalam statistika/matematika, LABEL/CLASS/TARGET.

Supervised dan Unsuoervised Learning

Machine Learning itu terbagi menjadi 2 tipe yaitu supervised dan unsupervised Learning. Jika LABEL/CLASS dari dataset sudah diketahui maka dikategorikan sebagai supervised learning, dan jika Label belum diketahui maka dikategorikan sebagai unsupervised learning.

Mengenali email sebagai spam atau bukan spam tergolong sebagai supervised learning, karena kita mengolah dataset yang berisi data point yang telah diberi LABEL "spam" dan “not spam”. Sedangkan jika kita ingin mengelompokkan customer ke dalam beberapa segmentasi berdasarkan variabel-variabel seperti pendapatan, umur, hobi, atau jenis pekerjaan, maka tergolong sebagai unsupervised learning

Algoritma Machine Learning

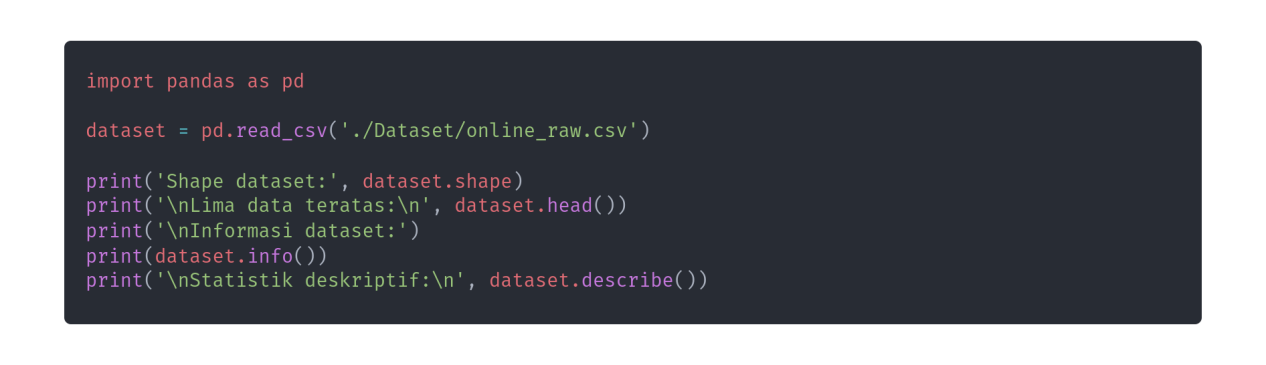
Penting untuk diingat bahwa tidak ada ML algorithm yang cocok atau fit untuk diaplikasikan di semua problem. Oleh karena itu, proses ini terkadang memerlukan trial & error seperti research, bahkan experienced data scientist pun tidak akan tahu apakah algorithm itu akan tepat atau tidak jika tidak mencoba.

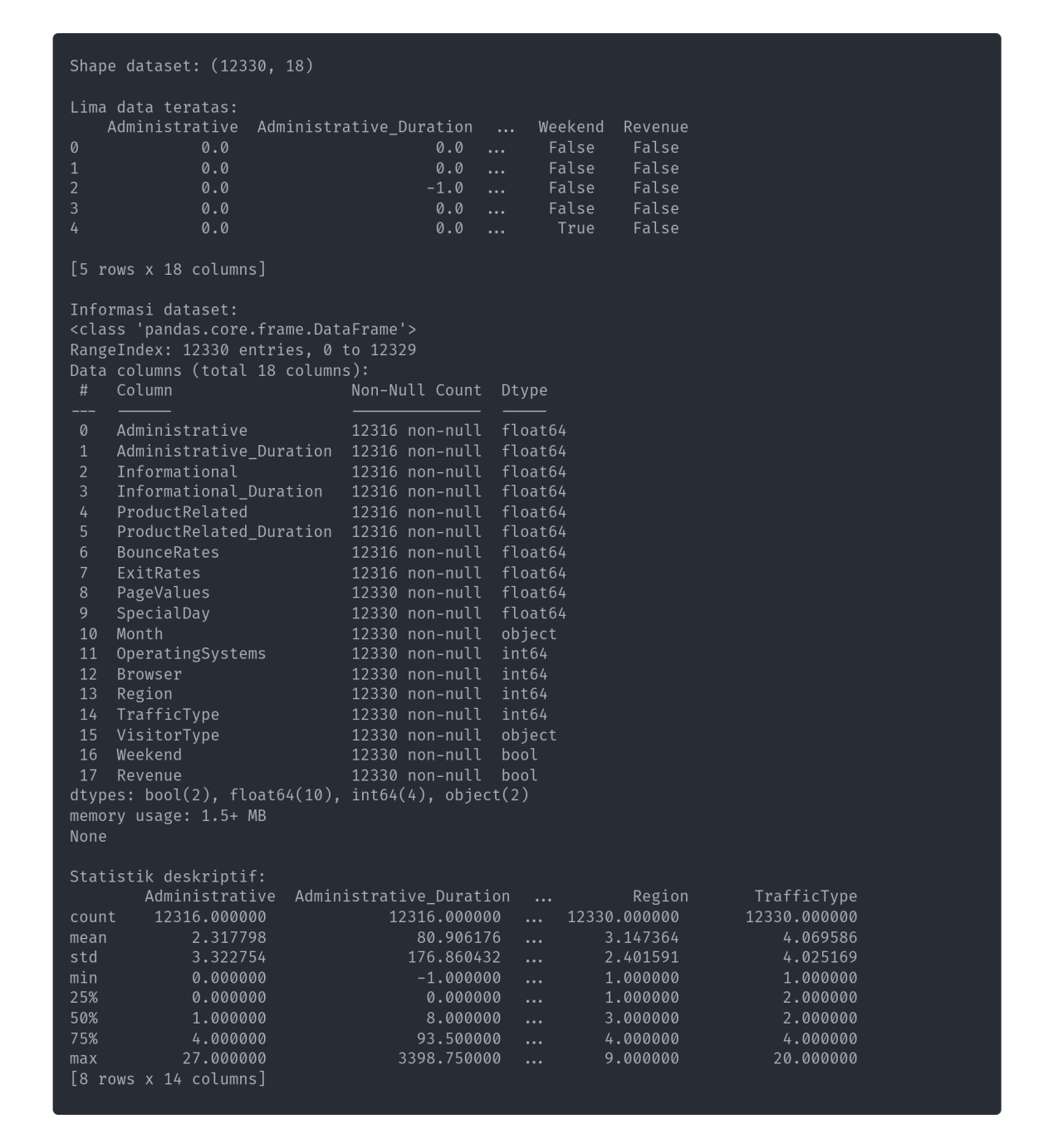
Algorithm dengan performansi yang paling baiklah yang dipilih sebagai model, selain itu untuk supervised learning, jika LABEL dari dataset kalian berupa numerik atau kontinu variabel seperti harga, dan jumlah penjualan, kita memilih metode REGRESI dan jika bukan numerik atau diskrit maka digunakan metode KLASIFIKASI. Untuk unsupervised learning, seperti segmentasi customer, kita menggunakan metode CLUSTERING

Eksplorasi Data & Data Pre-processing

Membuat model machine learning tidak serta-merta langsung modelling, ada tahapan sebelumnya yang penting untuk dilakukan sehingga kita menghasilkan model yang baik, Kita akan memanfaatkan Pandas library. Pandas cukup powerful untuk digunakan dalam menganalisa, memanipulasi dan membersihkan data.

Memahami Data Dengan Statistik

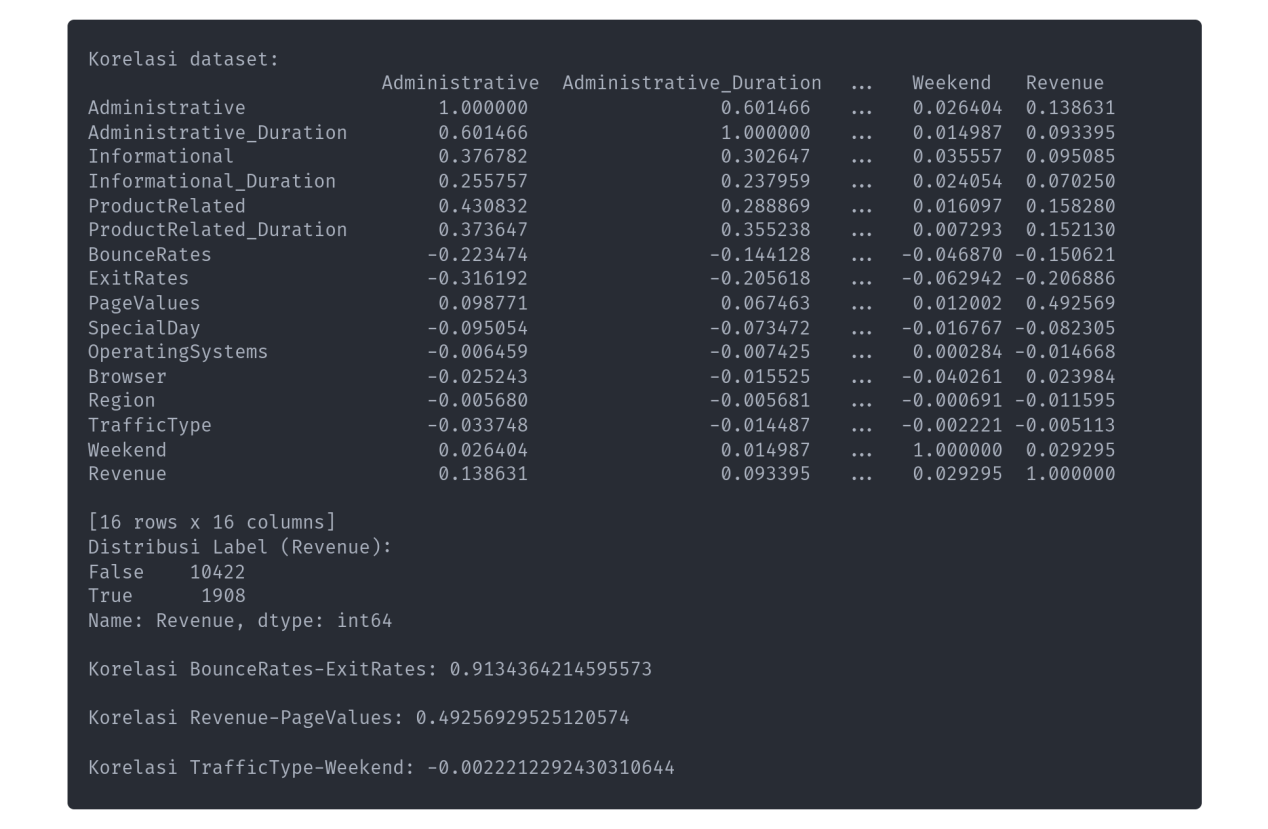




Dengan mengetahui dimensi data yaitu jumlah baris dan kolom kita dapat mengetahui apakah data yang kita miliki terlalu banyak atau terlalu sedikit. dimana semakin banyak data maka semakin lama pula proses pelatihan datanya nanti. dan jika data terlalu sedikit mungkin performa model tidak cukup baik.

selanjutnya eksplorasi data untuk melihat korelasi dan distribusi dataset



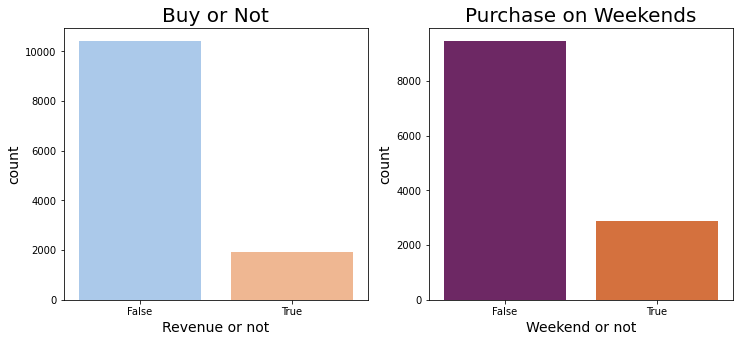


Mengetahui Label Distribusi Dataset itu Penting !

Mengetahui distribusi label sangat penting untuk permasalahan klasifikasi, karena jika distribusi label sangat tidak seimbang, maka akan sulit bagi model untuk mempelajari pola dari LABEL yang sedikit dan hasilnya bisa misleading

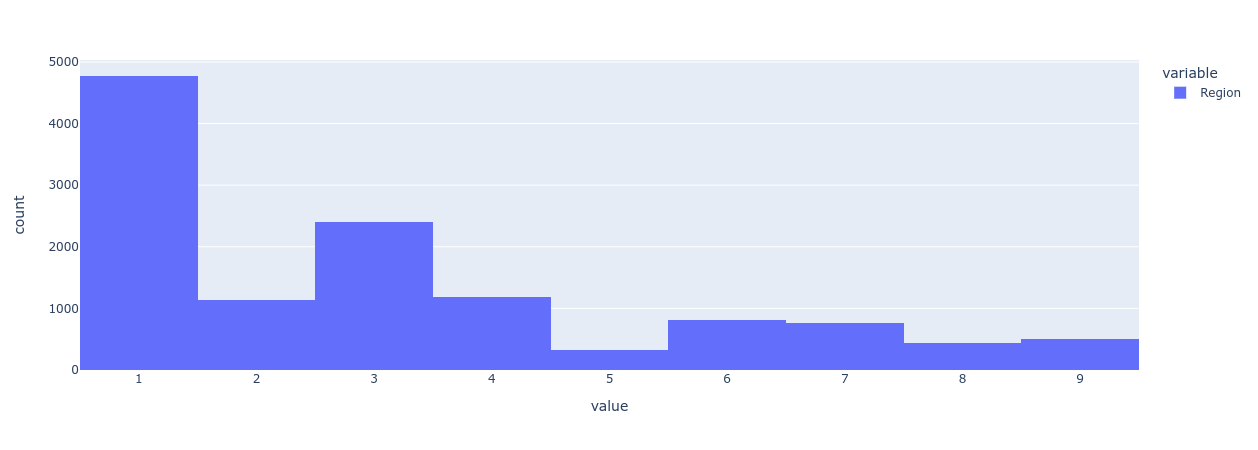
Dengan visualisasi kita dapat dengan mudah dan cepat dalam memahami data. Dengan membuat chart yang menampilkan perbandingan jumlah yang membeli (1) dan tidak membeli (0), serta perbandingan jumlah pembelian saat weekend.



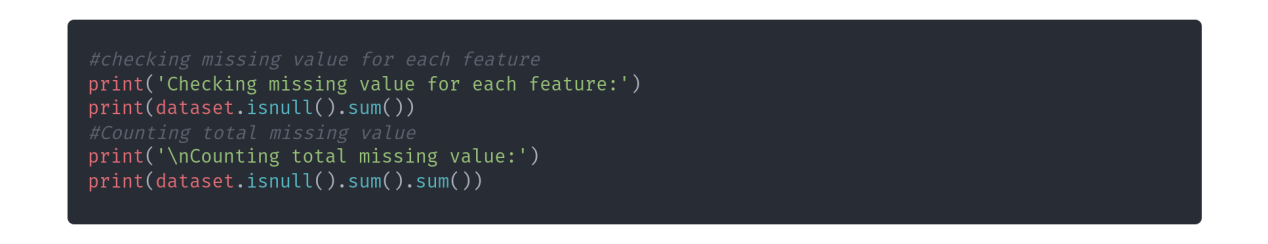


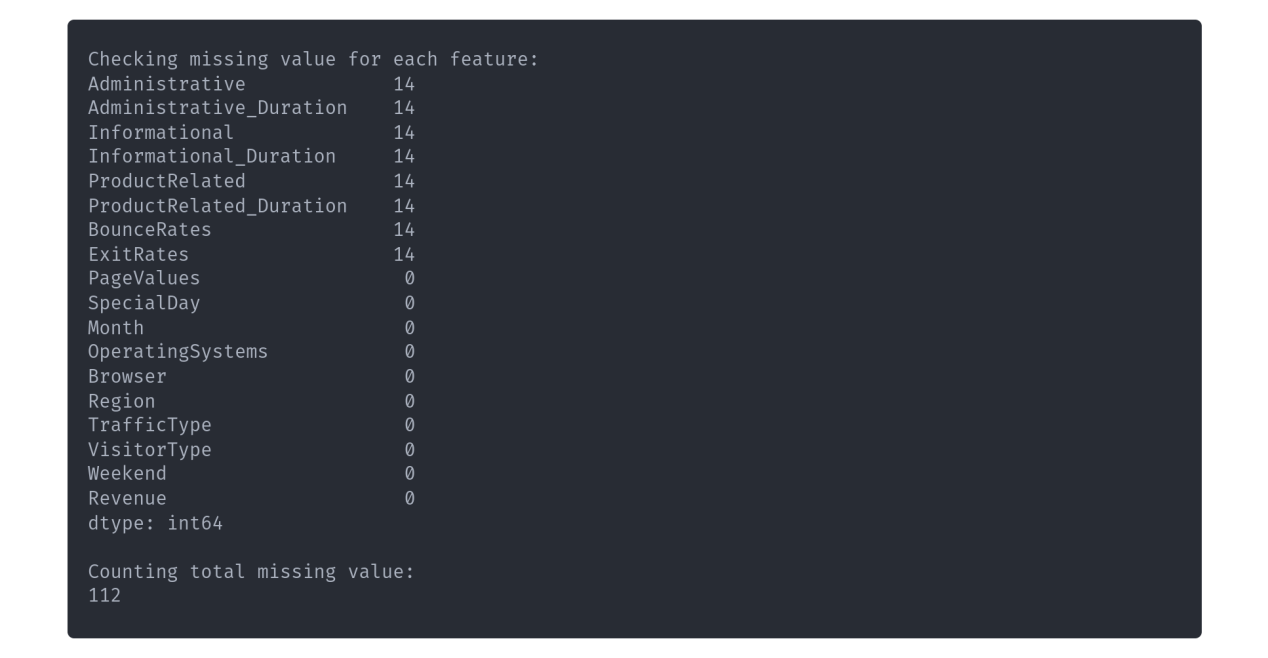
Membuat visualisasi data dengan histogram yang menggambarkan jumlah customer untuk setiap region.



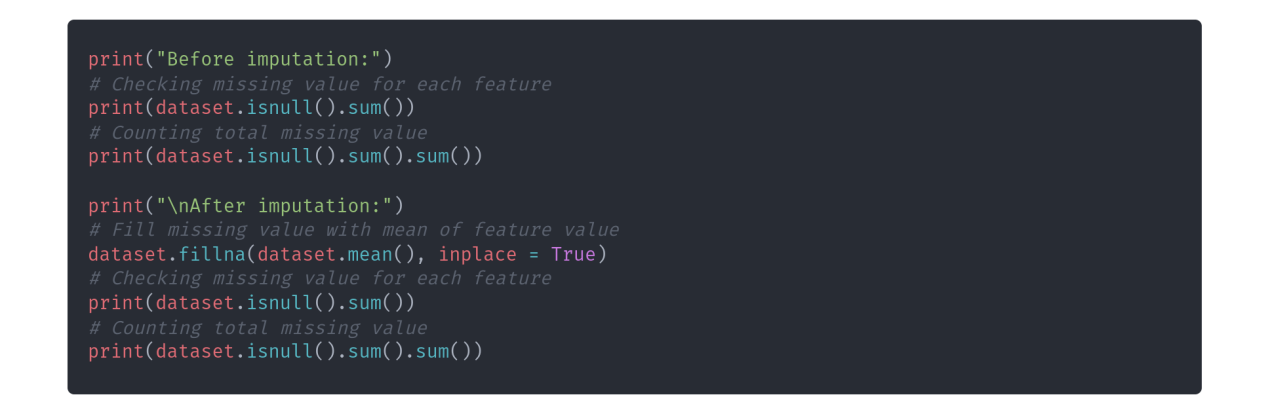


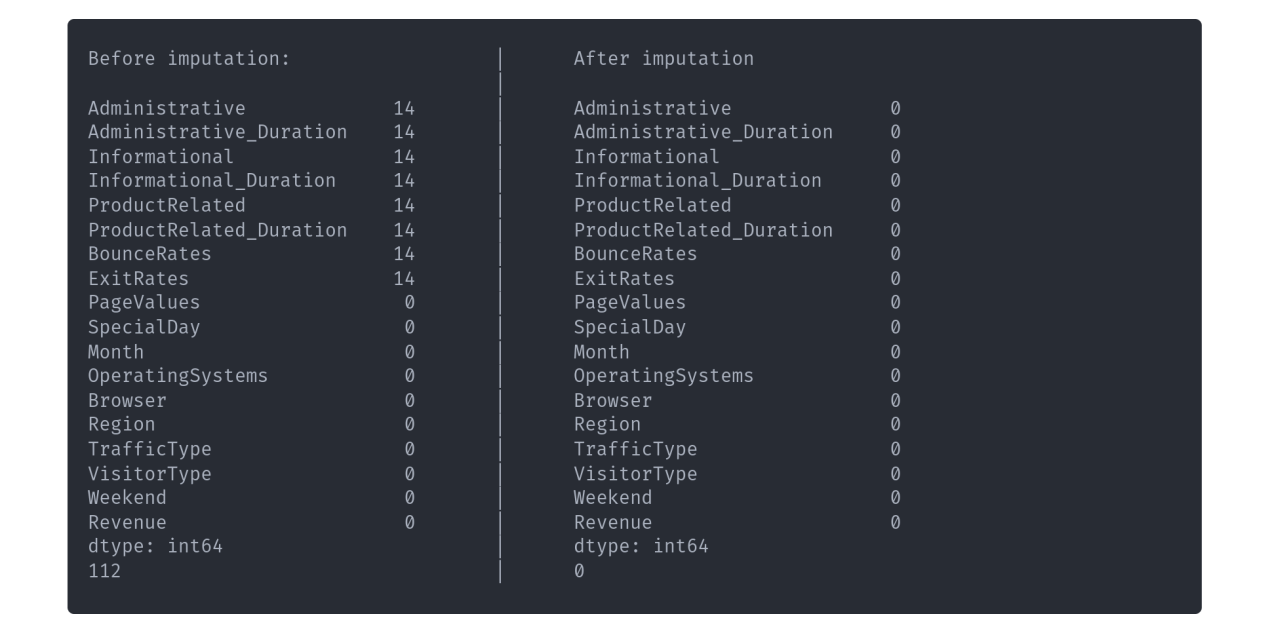
**Data Pre-processing: Handling Missing Value**



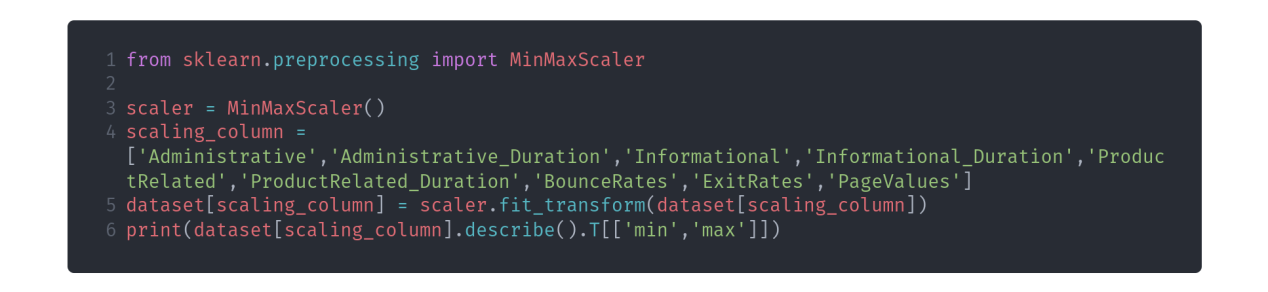


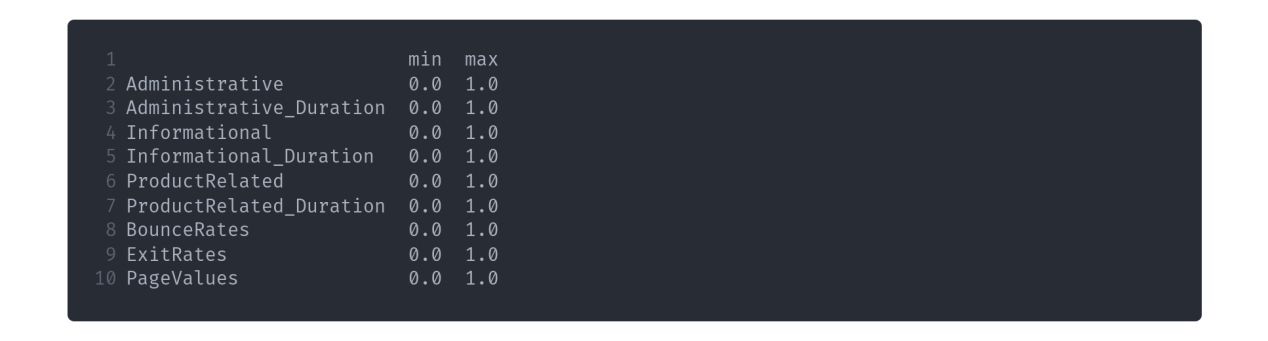
Imputing missing value sangat mudah dilakukan di Python, cukup memanfaatkan fungsi .fillna() dan .mean() dari Pandas, seperti berikut:



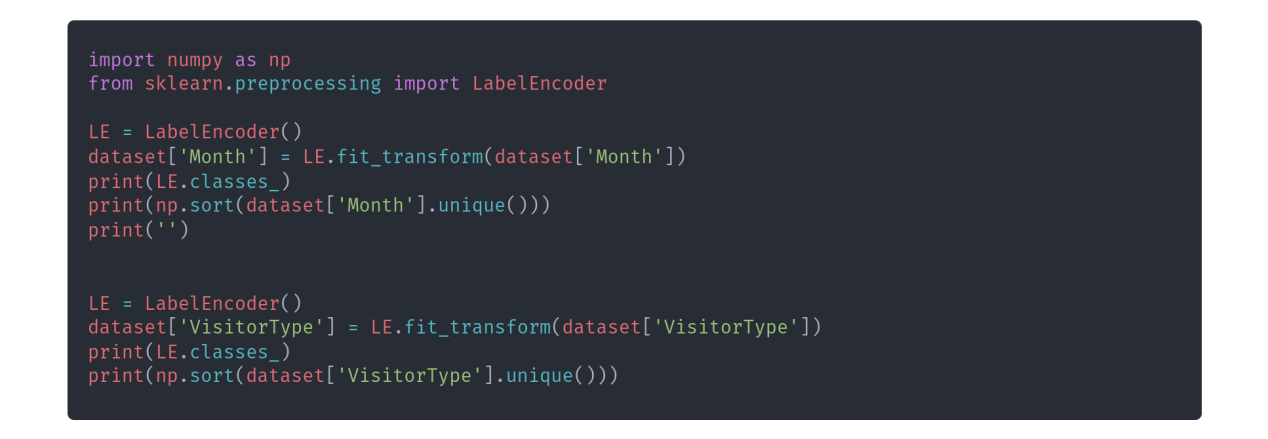


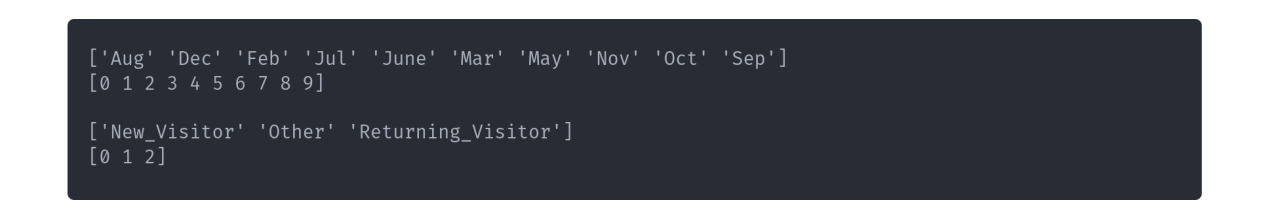
**Data Pre-processing: Scalling**





**Data Pre-processing: Konversi string ke numerik**

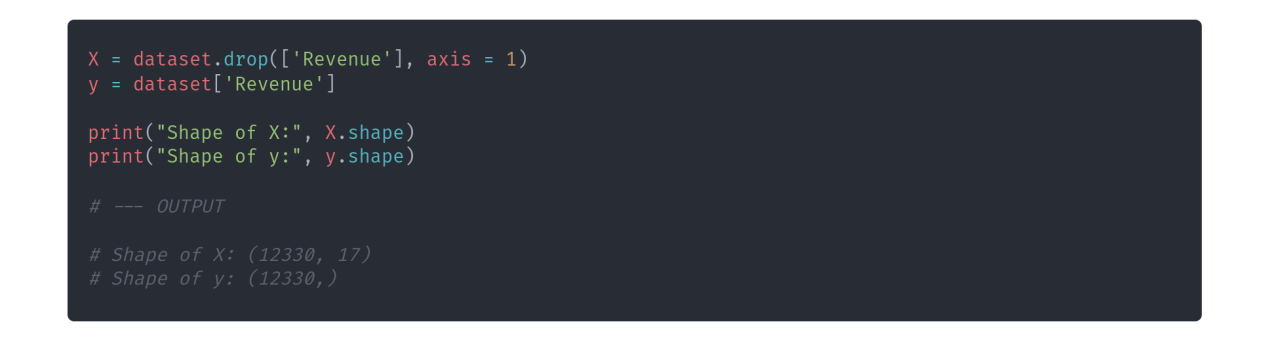




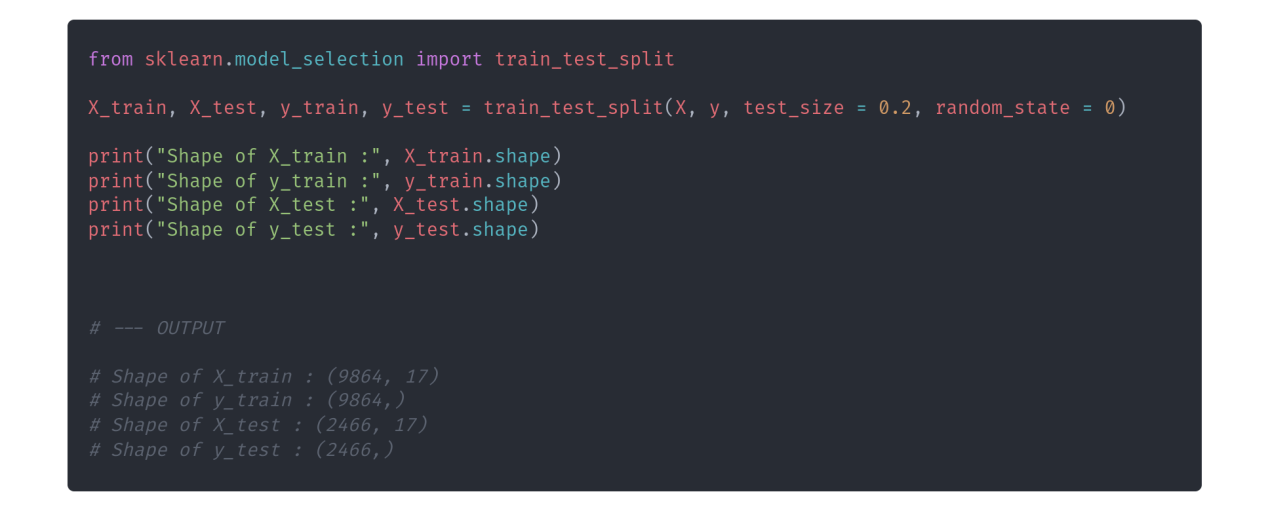
Pemodelan Dengan Scikit-Learn

**Features & label**

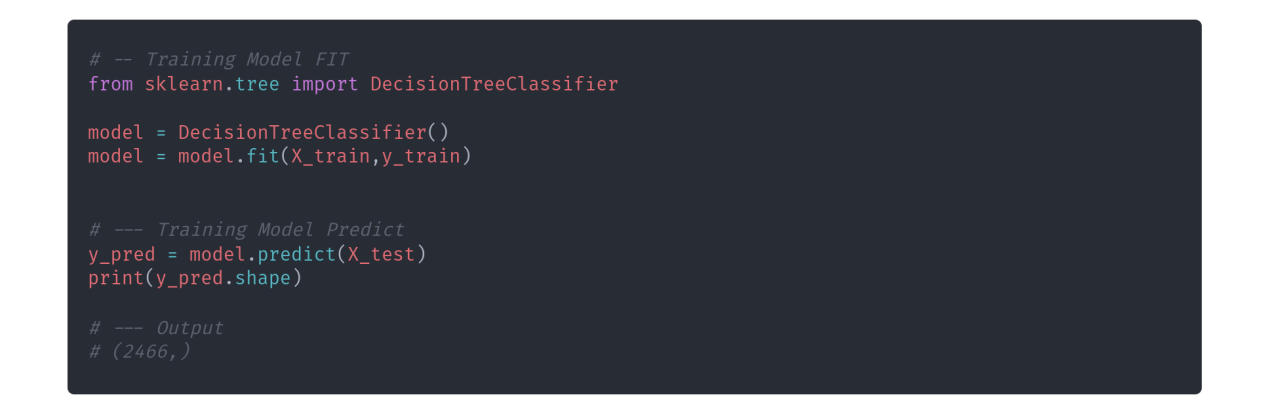
Dalam dataset user online purchase, label target sudah diketahui, yaitu kolom Revenue yang bernilai 1 untuk user yang membeli dan 0 untuk yang tidak membeli, sehingga pemodelan yang dilakukan ini adalah klasifikasi. Nah, untuk melatih dataset menggunakan Scikit-Learn library, dataset perlu dipisahkan ke dalam Features dan Label/Target. Variabel Feature akan terdiri dari variabel yang dideklarasikan sebagai X dan [Revenue] adalah variabel Target yang dideklarasikan sebagai y. Gunakan fungsi drop() untuk menghapus kolom [Revenue] dari dataset.



**Trainning Dan Test Dataset**



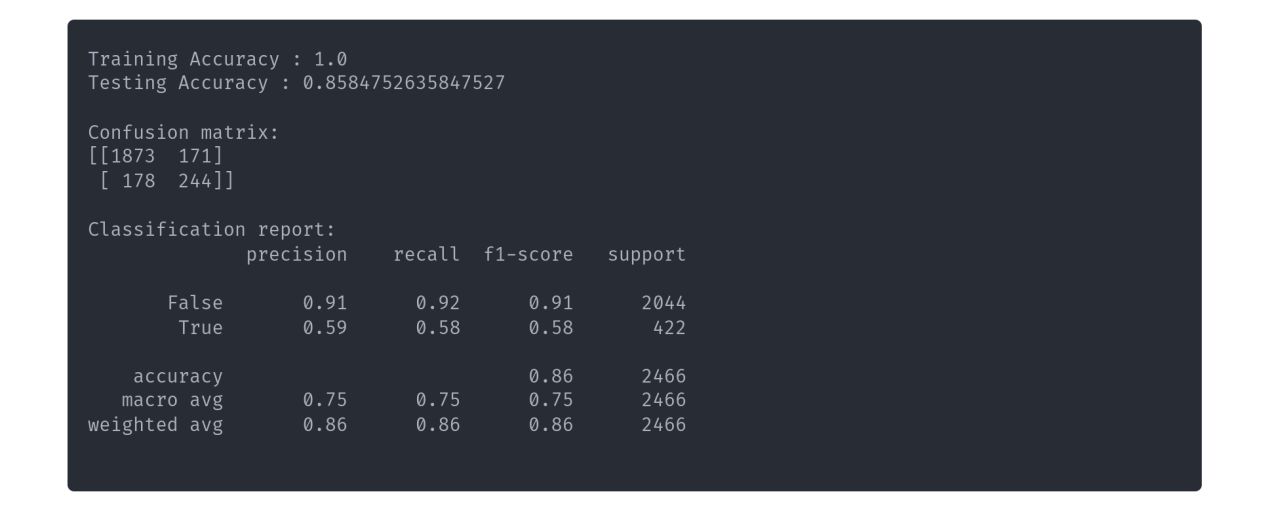
**Training Model: Fit & Predict**



**Evaluasi Model Performance**

Untuk menampilkan confusion matrix cukup menggunakan fungsi confusion\_matrix() dari Scikit-Learn.



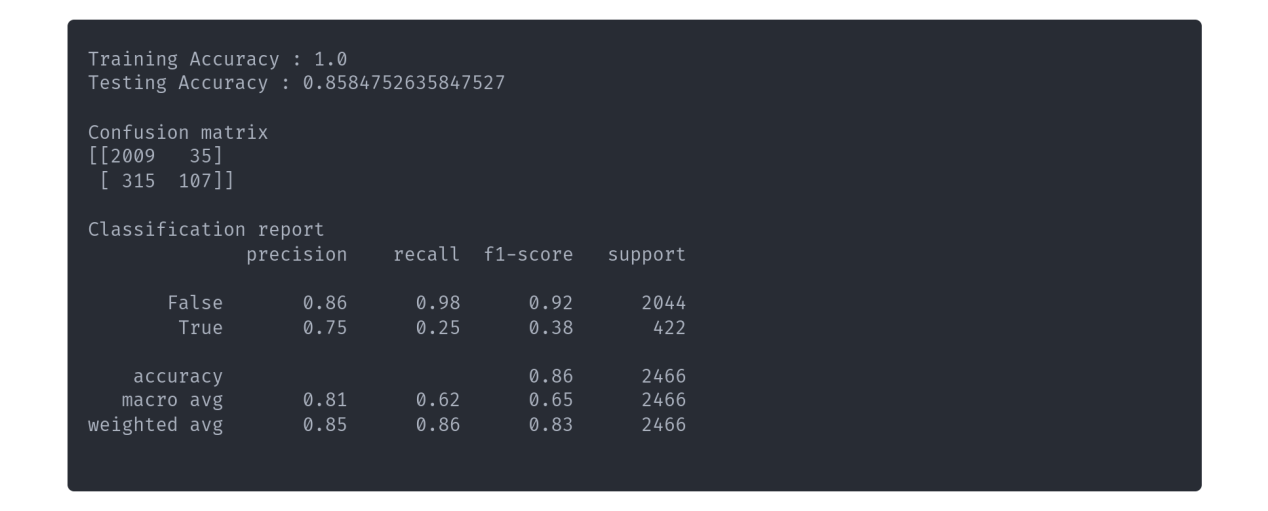


Pemodelan Permasalahan Klasifikasi dengan Logistic Regression

Pemodelan Logistic Regression dengan memanfaatkan Scikit-Learn sangatlah mudah. Dengan menggunakan dataset yang sama yaitu online\_raw, dan setelah dataset dibagi ke dalam Training Set dan Test Set, cukup menggunakan modul linear\_model dari Scikit-learn, dan memanggil fungsi LogisticRegression() yang diberi nama logreg.

Kemudian, model yang sudah ditraining ini bisa digunakan untuk memprediksi output/label dari test dataset sekaligus mengevaluasi model performance dengan fungsi score(), confusion\_matrix() dan classification\_report().





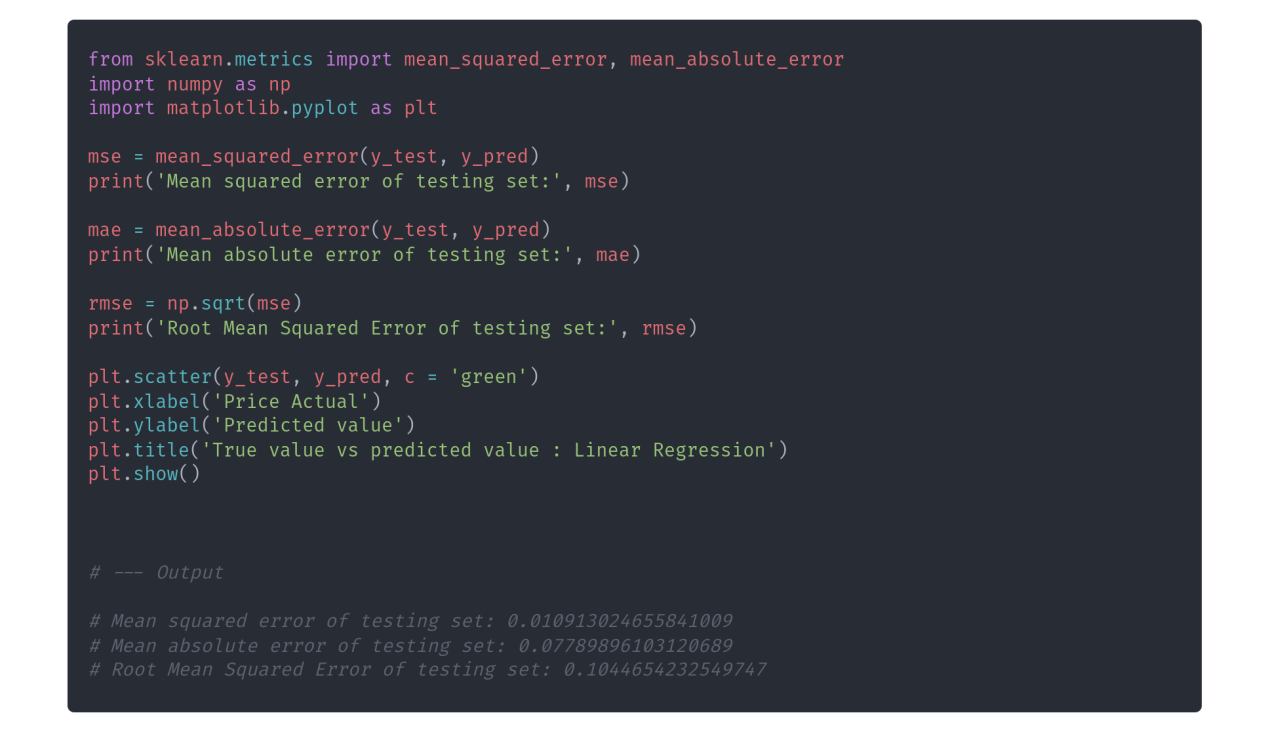
**Classification - Decision Tree**

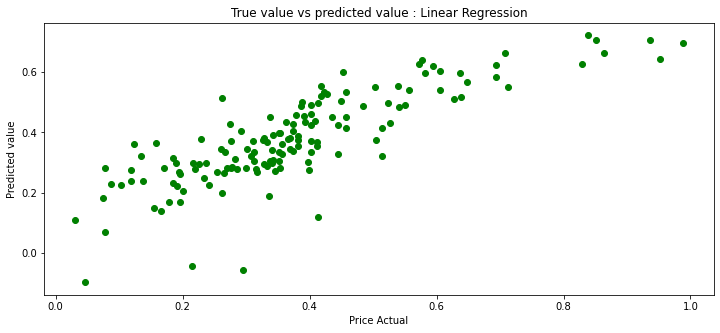


**Regression: Linear Regression**

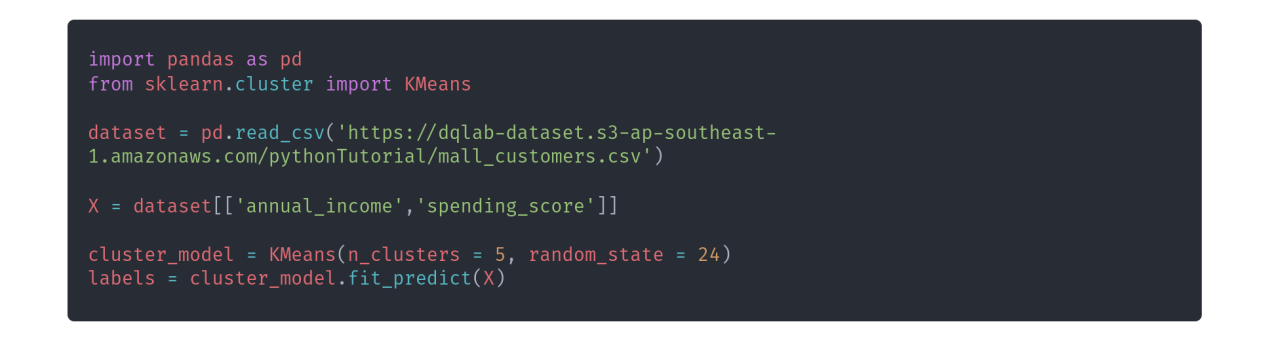


**Regression Performance Evaluation**

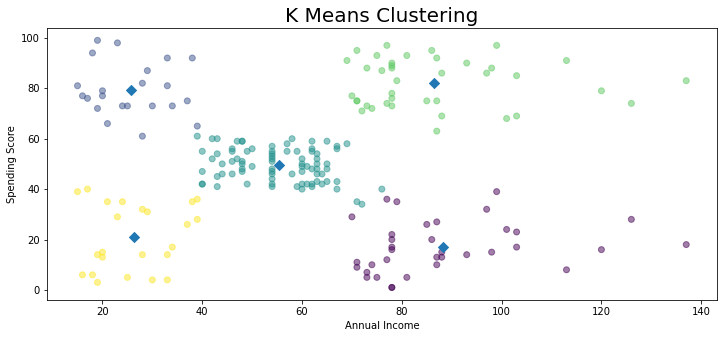




1. **Means Clustering**







**Measuring Cluster Criteria**

